Perceptual and model-based evaluation of ideal time-frequency noise reduction in hearing impaired listeners

Raphael Koning, Ian Bruce, Sam Denys, and Jan Wouters
E-mail: raphael.koning@med.kuleuven.be

***Abstract*—State-of-the-art hearing aids (HAs) try to overcome the deficit of poor speech intelligibility (SI) in noisy listening environments using digital noise reduction (NR) techniques. The application of time-frequency masks to the noisy sound input is a common NR technique to increase SI. The binary mask with its binary weights and the Wiener filter with continuous weights are representatives of a hard- and a soft-decision approach for time-frequency masking. In normal-hearing listeners, the ideal Wiener filter (IWF) outperforms the ideal binary mask (IBM) in terms of SI and speech quality with perfect SI even at very low signal-to-noise ratios (SNRs). In this study, both approaches were investigated for hearing-impaired (HI) listeners. Perceptual and auditory model-based measures were used for the evaluation. The IWF outperformed the IBM in terms of SI. Quality-wise, there was no overall difference between the NR algorithms perceived. Additionally, the processed signals were evaluated based on an auditory nerve model using the neurogram similarity metric (NSIM). The mean NSIM values were significantly different for intelligible and unintelligible sentences. The results suggest that a soft-mask seems to be promising for application in HAs.**

***Index Terms*—hearing aids, auditory prostheses, speech enhancement, noise reduction, time-frequency masking**

I. Introduction

ONE of the major challenges that people with sensorineural hearing loss face in their daily life is listening to a speaker in adverse listening conditions with an interfering background noise and/or reverberation. The speech reception threshold (SRT), which is the signal-to-noise ratio (SNR) required for 50% of the target speech material to be recognized correctly, for persons with hearing impairment can be easily $2-5$ dB higher in stationary speech-shaped noise than in normal-hearing (NH) listeners [1]. The increase in SRT for hearing-impaired (HI) subjects is even more prominent for fluctuating maskers and can exceed $7-15$ dB [2, 3, 4]. In all of these studies, the amplitude of the signal was adapted to the hearing level (HL) of the respective subject to prevent an influence of the loss of audibility on the results. There is more than just a loss of audibility and shift of audiometric thresholds accompanied with a sensorineural hearing loss. Most HI listeners suffer from a reduced frequency resolution [5], reduced masking release [6, 7], and limited use of acoustic temporal fine structure [8]. A lot of these deficits are based on broader auditory tuning. An exhaustive literature review on broader auditory tuning and its consequences can be found in [9].

To overcome the problems with speech perception in adverse listening conditions, noise reduction (NR) algorithms have been developed with the objective to improve speech intelligibility, speech quality and listening comfort. Most state-of-the art hearing aids (HAs) have a digital NR stage in their signal processing chain. The principle of NR is to reduce as much noise as possible from the noisy mixture under the constraint of limiting distortions of the target signal.

Single channel NR algorithms are often based on spectral subtraction [10], statistical modeling [11], or Wiener filtering [12]. In NH listeners, a comparison between eight noise reduction algorithms revealed that none was able to improve speech intelligibility (SI) in various noisy conditions [13] but some were able to enhance speech quality [14]. In general, speech quality improvements were observed for single channel noise reduction algorithms [13, 15, 16, 17, 18], but they were unable to increase SI [14, 18, 19, 20].

Another single channel NR approach is to apply a time-frequency mask to the spectrum of the noisy signal with gains that are dependent on the SNR in the respective time-frequency point. A literature review on time-frequency masking can be found in [21, 22]. There are two very prominent approaches based on administering either binary gains, leading to a so-called binary mask, or continuous weights between 0 and 1, leading to a so-called soft mask.

The approach to administer binary values is motivated by the auditory masking phenomenon. With its binary values, the binary mask (BM) exploits the sparsity and disjointness of the target and interferer spectra. Under ideal parameter conditions, often referred to as ideal binary mask (IBM), the mask preserves time-frequency points where the SNR is above a certain threshold value and suppresses completely the other remaining time-frequency points. The IBM was suggested to be the target goal of computational auditory scene analysis [23] and provides with a threshold value of 0 dB the optimal SNR gain of all binary masks [24]. It was shown that approaches based on binary masks with and without *a priori* knowledge for the mask computation can increase under certain conditions SI in NH [25, 26, 27, 28, 29, 30] and HI listeners [26, 28, 30].

Speech quality comparisons, however, revealed that the so-called soft decision approaches with masks consisting of continuous weights between 0 and 1 outperformed the binary decision approaches in terms of speech quality [13]. The most popular representative of the masks with continuous weights is the Wiener filter. Under ideal parameter knowledge for the mask pattern derivation, it is often referred to as ideal Wiener filter (IWF).

A comparison between the two approaches in [31] revealed that in terms of potential for SI and speech quality improvement for NH listeners, the IWF vastly outperformed the IBM with a frequency resolution according to the Bark-scale. The coarse spectral resolution of the Bark-scale is close to or even higher than the frequency resolution that NR algorithms operate on in state-of-the-art HA. Furthermore, it was shown that the SI of the soft mask processed signals was more robust compared to the binary mask processed signals when estimation errors in the applied mask were simulated. It was concluded in [31] that the IWF approach should be preferred over the IBM approach in auditory prostheses such as HAs and cochlear implants (CIs). A CI is an auditory prostheses for people with profound sensorineural hearing loss that evokes an auditory sensation by electrical stimulation of the auditory nerve. In CI users, the NR algorithms did not differ significantly in terms of SI improvement and robustness to estimation errors [32]. The choice of the NR algorithm between IBM and IWF was not important for applications in CIs. Therefore, it remains unclear what the potential and the limits of both time-frequency masks are in HI listeners.

The primary goal of this study is to investigate the potential of the IWF and IBM for its application in HAs with emphasis on SI and

The work of R. Koning was supported by the EU within the Marie Curie ITN AUDIS, Grant Agreement No. PITN-GA-2008-214699. The work of I. Bruce was supported by the Natural Sciences and Engineering Research Council of Canada, Discovery Grand 2617636.

speech quality improvement. It is investigated whether the SI is dependent on the degree of hearing loss of the subject. Furthermore, the robustness of the SI is studied when estimation errors are simulated in the mask pattern. The results of the speech recognition tasks of NH and HI listeners under the assumption of ideal parameter estimates are evaluated with the auditory-periphery model of [33], and using the instrumental measure neurogram similarity metric (NSIM). The NSIM tries to predict the SI by comparing the neurogram simulated for the hearing loss of the respective subject with a neurogram that is obtained for NH listeners. NSIM was shown to be accurate for different presentation levels of speech in quiet for various HLs.

II. Signal processing

The NR signal processing was described in detail in [31]. The additive signal model of a target speech signal $s(t)$ and an interfering sound $v(t)$ resulting in the microphone signal $y(t)$ can be written in the short-time frequency domain with the frame index n and the frequency index k as

$$Y(n,k) = S(n,k) + V(n,k). \tag{II.1}$$

The aim of applying a time-frequency mask $G(n,k)$ to the microphone signal is to obtain an estimate $\hat{S}(n,k)$ of the target speech signal. The output of the NR step can be written as

$$\hat{S}(n,k) = G(n,k)\, Y(n,k). \tag{II.2}$$

Both approaches investigated in this study derive their gain function as a function of the short-term SNR $\xi(n,k)$, which is defined as the ratio between the power spectral density of the target signal $\Phi_{SS}(n,k)$ and the interfering sound $\Phi_{VV}(n,k)$. Under the assumption of perfect knowledge of both components, the power spectral densities can be substituted by the instantaneous powers of the respective signals. The short-term SNR can be written as

$$\xi(n,k) = \frac{\Phi_{SS}(n,k)}{\Phi_{VV}(n,k)} = \frac{|S(n,k)|^2}{|V(n,k)|^2}. \tag{II.3}$$

In this study, *a priori* knowledge of the target and interfering background sound is used to derive the gain factors of the mask $G(n,k)$.

A. Ideal binary mask (IBM)

In this study, an IBM with a local threshold [31, 34] is used, where a gain of 1 is accorded to the mask when the short-term SNR is above the global input SNR ξ_{in} of the overall mixture of the speech and the interfering sound. The binary mask G_{IBM} can be written as

$$G_{\mathrm{IBM}}(n,k) = \begin{cases} 1 & \text{if } \xi(n,k) > \xi_{\mathrm{in}}, \\ 0 & \text{else.} \end{cases} \tag{II.4}$$

B. Ideal Wiener filter

In contrast to the IBM with its binary weights, the mask of the IWF G_{IWF} applies continuous values between 0 and 1 and can be written as

$$G_{\mathrm{IWF}}(n,k) = \frac{\xi(n,k)}{1+\xi(n,k)}. \tag{II.5}$$

(II.5) is obtained as the minimum mean square error (MMSE) estimate of the complex spectral amplitude [21].

The mask patterns in (II.4) and (II.5) are then applied to the noisy mixture in (II.2) to obtain the processed signal.

C. Simulation of estimation errors

Over-and underestimation errors of parameters used in NR systems influence speech intelligibility differently [35]. To ensure the equal amount of under- and overestimation of the instantaneous power spectral density that are used to calculate the short-term SNR in (II.3), an additional white noise ϵ term with zero mean and power equal to the clean target signal was added to the spectrum of the target and the interfering signal in each frequency band [31]. The corrupted spectrum can be written as

$$\tilde{S}(n,k) = S(n,k) + \epsilon_S(n,k) \tag{II.6}$$
$$\tilde{V}(n,k) = V(n,k) + \epsilon_V(n,k). \tag{II.7}$$

The corrupted spectra perturb both mask patterns over the instantaneous SNR in (II.3). To avoid confusion with the mask pattern with ideal estimates, the conditions with perturbed parameter estimates are called BM and Wiener filter (WF) with the respective masks G_{BM} and G_{WF}. The masks of the BM and WF are then applied to the noisy mixture of the speech and interfering background signal in (II.2).

D. General processing steps

All stimuli were sampled at a sampling rate of $f_s = 16000$ Hz and were transformed in the frequency domain with a Fast Fourier Transform (FFT) of 512 points resulting in a frame length of 32 ms. Furthermore, a frame-shift of 16 ms was applied. The signal processing in HAs and CIs is not based on such a high resolution. Therefore, a frequency resolution was chosen as in [31] according to the Bark scale [36]. Due to the sampling rate of $f_s = 16000$ Hz the is done in 22 instead of 24 critical bands of hearing as the 2 bands with the highest frequency content up to 15500 Hz cannot be resolved with this sampling rate. To obtain this coarse spectral resolution, an analysis window (square-root Hann window of a 32 ms length) was applied to each frame to compute the FFT coefficients. Afterwards, the magnitude-squared coefficients were grouped according to the Bark-scale and the instantaneous SNR in (II.3) was computed. The instantaneous SNR was then used to calculate the gain function according to (II.4) and (II.5) and applied to the mixture signal spectrum. After the noise reduction stage, the transformation to the time domain was done by applying an inverse FFT to the frame weighted with a synthesis windows. The synthesis window was again a square-root Hann-window of the frame length. For the HI listeners, an additional amplification stage was introduced according to the NAL-RP rule [37] to compensate for the hearing loss of the respective subject.

III. Methods

A. Subjects

Two groups of listeners participated in the speech recognition tasks: 6 NH and 9 HI listeners. All NH subjects had hearing thresholds better than 20 dB HL at the octave frequencies between 125 Hz and 8000 Hz. The mean age of this group was 21 years (with a standard deviation of 2 years). They were not paid for their participation. The group of HI listeners had an average of 67 years (with a standard deviation of 10 years). The audiometric thresholds, test ear, and the pure tone average (PTA) at frequencies of 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz of all HI subjects are shown in Table I. Travel expenses of the HI listeners were reimbursed. All subjects were native Dutch speakers. They signed an informed consent form before the tests were conducted.

TABLE I
DETAILS OF THE PARTICIPATING HI SUBJECTS: AGE (YEARS), TESTED EAR, THE RESPECTIVE AUDIOMETRIC THRESHOLDS (DB HL), AND THE PTA

			Audiometric threshold [dB HL] Frequency [kHz]							
Sub.	Age	Ear	0.125	0.25	0.5	1	2	4	8	PTA
1	57	L	15	15	15	20	60	55	50	38
2	79	R	10	10	30	25	60	75	95	48
3	75	R	25	25	30	40	65	75	75	53
4	58	L	30	40	50	55	70	80	90	64
5	51	L	5	5	20	35	35	60	65	38
6	62	R	15	5	10	10	20	40	60	20
7	76	L	5	15	15	25	40	65	80	36
8	78	R	15	20	25	35	60	65	75	46
9	68	R	10	5	20	20	50	60	70	38

TABLE II
CONDITIONS OF THE SENTENCE RECOGNITION TASKS

Processing	Noise type	Speech material	Mixing SNR ξ_0 [dB]
IWF/IBM	Babble	VU/LIST	-30,-20,...,0
IWF/IBM	Speech	VU	-30,-20,...,0
WF/BM	Babble	VU/LIST	0

B. Testmaterial

Two Dutch speech corpora were used as the target sentences in the speech recognition tasks: the University Hospital VU, Amsterdam, sentences [38] and the Leuven Intelligibility Sentence Test (LIST) sentences [39].

A female speaker of the Dutch VU sentences was chosen as the target speech material. The VU sentences speech material consists of 39 lists of 13 sentences. The speaking rate is 4.7 syllables per second which is representative for a conversation. Scoring was done on sentence level, i.e., a score of 100 % was given if all the words in a sentence were reported correctly by the subject, whereas a score of 0 % was given if one or more words in the sentence were reported incorrectly by the subject. A male speaker of the VU sentences was selected as the interfering speaker in the speech-in-speech scenario.

Additionally to the VU sentences, the listening tasks were also performed with the LIST sentences. The LIST consists of 35 lists of 10 sentences of a female target speaker. Per list, there are 32 to 33 keywords assigned which allow scoring on the keyword and sentence level. The overall speaking rate of the LIST sentences is 2.5 syllables per second. The LIST sentences were developed as a target speech material for cochlear implant users or HI subjects with severe hearing loss. Therefore, the speech rate is about half the rate of the VU sentences.

The Auditec multi-talker babble noise (from the CD *Auditory Tests (Revised), Auditec, St. Louis, MO, USA*) was used as the interfering background sound in the speech-in-babble scenario.

C. Procedure of the perceptual evaluation

Both groups of listeners participated in sentence recognition tasks with ideal mask estimates. The HI listeners participated additionally in a sentence recognition task with perturbed mask estimates and a quality rating. A test-retest design was applied to evaluate both NR algorithms with regard to the test-retest reliability and learning effects. There was at least one week between the two sessions. Each session lasted around two hours. All tests were performed double-blind – the subject and the experimenter did not know the condition of the respective trial. To avoid effects based on listening fatigue, subjects were allowed to take breaks during the testing. In all listening tasks, the level of the clean speech signal was set to $65\,\mathrm{dB}$ sound pressure level (SPL). All signals were rescaled to the level of the clean speech signal to prevent audibility issues at low SNRs. In a sound-proofed booth, The stimuli were presented monourally with a Sennheiser HDA200 headphone on the left or the better ear for the NH and HI subjects, respectively. For the HI listeners, the better ear was defined as the ear with the lower pure tone average. An additional amplification set according to the NAL-RP rule [37] was introduced to compensate for the subject's HL.

1) Speech intelligibility with ideal parameter estimates: The mask patterns for the IBM and the IWF were calculated with ideal parameter knowledge as in (II.4) and (II.5). The signals were mixed at SNRs of $0\,\mathrm{dB}$ to $-30\,\mathrm{dB}$ with a stepsize of $-10\,\mathrm{dB}$. At each SNR, the scores were determined based on one list of the respective speech material. The order of algorithm, noise scenario and SNR was randomized. All combinations were tested with each subject.

2) Robustness to estimation errors: The second listening task was performed with perturbed mask estimates ((II.6) and (II.7)) at an SNR of $\xi_{\mathrm{in}} = 0\,\mathrm{dB}$ in the speech-in-babble scenario with both speech materials.

A summary about the conditions that were tested in the speech recognition tasks can be found in Table II.

3) Preference rating: The perceived quality was studied with a two-stage pairwise preference rating test [18, 31]. The pairwise comparisons were conducted across

i) clean speech and IWF processed signal,
ii) clean speech and IBM processed signal,
iii) and IWF processed signal and IBM processed signal.

Both interfering scenarios (speech-in-speech and speech-in-babble) were tested at an SNR of $0\,\mathrm{dB}$ with the VU sentences, while the processed signals of the LIST sentences were tested in the speech-in-babble scenario. For each pairwise comparison, both signals were presented one after the other. The order was randomized. The subjects were allowed to repeat each individual stimulus as often as they wanted to.

The first stage of the two-stage procedure was an overall preference rating where the subject had to indicate which stimulus was preferred in terms of quality. In the next stage, they rated their preference on a 5 point scale ranging from *imperceptible* to *hugely better*. For each noise type, 10 sentences were used from one list of the VU sentences. The first 5 sentences were used during the test session, while the next five sentences were used in the retest session, in total 60 comparisons (10 sentences $\times$ 3 pairwise comparisons $\times$ 2 noise conditions). For the HI listeners, the quality rating was also conducted with one list of the LIST sentences in the speech-in-babble scenario.

D. Procedure of the model-based evaluation

In the model-based evaluation, an attempt was made to evaluate the processing strategies with an auditory-periphery model and to derive an objective measure to estimate the speech intelligibility.

An auditory-periphery model [33, 40, 41] was used that takes the effect of an impairment of the inner and the outer hair cells into account. It has been validated against a wide range of physiological data for both simple and complex stimuli. The model is able to generate a so-called neurogram in which the spike activity of a number of auditory nerve fibers is simulated over time. In this study, neurograms were calculated for 30 center frequencies with each fifty model auditory nerve fibers consisting of a physiologically-realistic mix of spontaneous rates and corresponding thresholds [42] with a time step of $6.4\,\mathrm{ms}$.

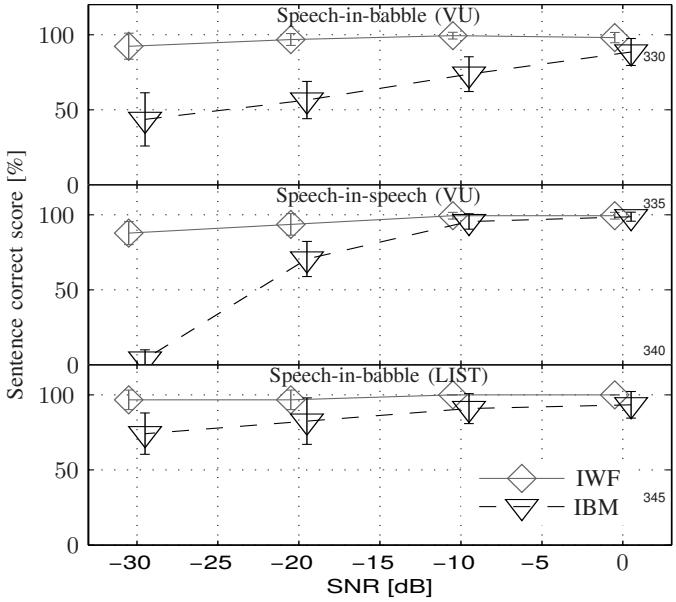

Fig. 1. Results of the 6 NH listeners in the sentence recognition task with ideal parameter estimates. The mean sentence correct scores for all subjects for the IWF and the IBM processed signals are marked by diamonds and triangles, respectively. The error bars depict one standard deviation of the scores.

For an average HI listener, the threshold shift in the audiogram can be attributed by two-thirds to an outer hair cell impairment and by one-third to an inner hair cell impairment [43]. Although some variation in the pattern of hair cell impairment across subjects can be expected [44, 45], a mixed impairment of outer and inner hair cells is likely the most prominent.

As an instrumental measure that is derived from the neurogram representation of the processed stimuli, the NSIM was used [46]. It is based on the structural similarity metric (SSIM) [47] which can predict the perceived quality of an image better than other pixel-by-pixel mean-squared error metrics. The NSIM was also validated for speech in quiet against different presentation levels and different degrees of HL. As the template neurogram, the neurogram of the clean speech presented at 65 dB SPL to a NH listener without HL was calculated with the same frequency and time-resolution of the test neurogram.

IV. Results of the perceptual evaluation

All percentage correct scores of the sentence recognition tasks were transformed to 'rationalized' arcsine units [48] before statistical analysis. Furthermore, the significance level of $p = 0.05$ was corrected with a Bonferroni correction for all multiple comparisons.

A. Speech intelligibility with ideal parameter estimates

Normal-hearing listeners: The results of the NH listeners in the sentence recognition task with ideal pattern estimates are shown for the IWF and the IBM processed signals in Fig. 1. The sentence correct scores with the IWF and the IBM processed signal are represented by the solid grey and the dashed black line, respectively. The error bars represent one standard deviation of the sentence correct scores. In Fig. 1, the sentence correct scores are shown for the VU sentences in the first two rows for the speech-in-babble (top) and the speech-in-speech (middle) scenario. The bottom row represents the sentence recognition scores in the speech-in-babble scenario when the LIST sentences served as the target speech material.

For the VU sentences, the results were evaluated with a four-way repeated measures analysis of variance (RM-ANOVA) with the factors NR algorithm, noise scenario, session and SNR. Mauchley's test of sphericity was passed for all four factors. The RM-ANOVA revealed significant effects of the main factors NR algorithm $[F(1,5) = 1186.5;\ p < 0.001]$ and SNR $[F(3,15) = 115.5;\ p < 0.001]$. The effect of the NR algorithm corresponded with a mean difference of $\Delta_{\mathrm{SI}} = \mathrm{SI}_{\mathrm{(I)WF}} - \mathrm{SI}_{\mathrm{(I)BM}} = 29.5\%$ in sentence intelligibility. There was no session effect obtained. Significant interaction effects were obtained between the factors noise scenario and SNR $[F(3,15) = 33.1;\ p < 0.001]$ and NR algorithm and SNR $[F(3,15) = 87.8;\ p < 0.001]$. A three factor interaction was obtained between the factors noise scenario, SNR and NR algorithm $[F(3,15) = 11.4;\ p < 0.001]$.

Post-hoc analysis of the sentence recognition task when the multitalker babble noise served as the interfering sound revealed that the SI between the IWF and the IBM processed signals was significantly different at 0 dB ($p < 0.05$), -10 dB ($p < 0.001$), -20 dB ($p < 0.05$), and -30 dB ($p < 0.001$). These significant differences corresponded to SI differences of $\Delta_{\mathrm{SI}} = 9.6\%$, $\Delta_{\mathrm{SI}} = 25.6\%$, $\Delta_{\mathrm{SI}} = 40.4\%$, and $\Delta_{\mathrm{SI}} = 48.7\%$, respectively. The decreasing SNR had no effect on the SI with respect to the IWF processed signals, while the sentence recognition scores for the IBM processed signals varied significantly ($p < 0.001$) with the SNR. In the speech-in-speech scenario, *post-hoc* analysis of the data revealed that the IWF and IBM processed signals differed significantly at SNRs of -20 dB ($p < 0.05$) and -30 dB ($p < 0.001$). The corresponding SI difference was $\Delta_{\mathrm{SI}} = 23.1\%$ and $\Delta_{\mathrm{SI}} = 84\%$, respectively. In contrast to the results in speech-in-babble, the input SNR had an influence on the IWF and the IBM ($p < 0.001$) processed signals.

To evaluate the SI scores obtained with the female speaker of the LIST sentences, a three-way RM-ANOVA with the factors NR algorithm, session and SNR was conducted. Mauchly's test of sphericity revealed no significant effects. No significant session effect was obtained. The main effects SNR $[F(3,15) = 8.7;\ p < 0.05]$ and NR algorithm $[F(1,5) = 55.0;\ p < 0.001]$ were significant. With the LIST speech material, the mean intelligibility difference was $\Delta_{\mathrm{SI}} = 13.1\%$. An interaction between the factors session and SNR $[F(1,5) = 6.9;\ p < 0.05]$ occurred. The SI with the IWF and IBM mask patterns differed significantly by $\Delta_{\mathrm{SI}} = 9.2\%$ at -10 dB ($p < 0.001$), $\Delta_{\mathrm{SI}} = 14.2\%$ at -20 dB ($p < 0.001$), and $\Delta_{\mathrm{SI}} = 22.5\%$ at -30 dB ($p < 0.001$) SNR. *Post-hoc* analysis revealed that the interaction effect of the SNR and session occurred at -20 dB SNR where the mean SI was significantly higher ($p < 0.001$) by 3.3% in the retest than in the test session.

Hearing-impaired listeners: The sentence recognition scores across sessions of the sentence recognition task with ideal parameter estimates are shown per subject in Fig. 2 for the VU sentences speech material. The sentence recognition scores with the IWF processed signals are shown with the grey solid line marked with empty and filled diamonds for the speech-in-babble and speech-in-speech condition, respectively. For the IBM processed signals, the scores are shown with the black dashed line with empty and filled triangles for the respective noise scenarios. The mean sentence correct scores of the group of HI listeners with ideal parameter estimates are shown in Fig. 3. As in Fig. 1 for the top and middle panels, the results for the VU sentences are shown in the speech-in-babble and the speech-in-speech scenario. In the bottom panel, the results for the LIST sentences in the speech-in-babble scenario are shown. The IWF condition is represented by the grey solid line marked with diamonds, while the IBM is represented by the black dashed line marked with triangles. The error bars depict one standard deviation of the scores.

The results with the VU sentences were evaluated with a four-way

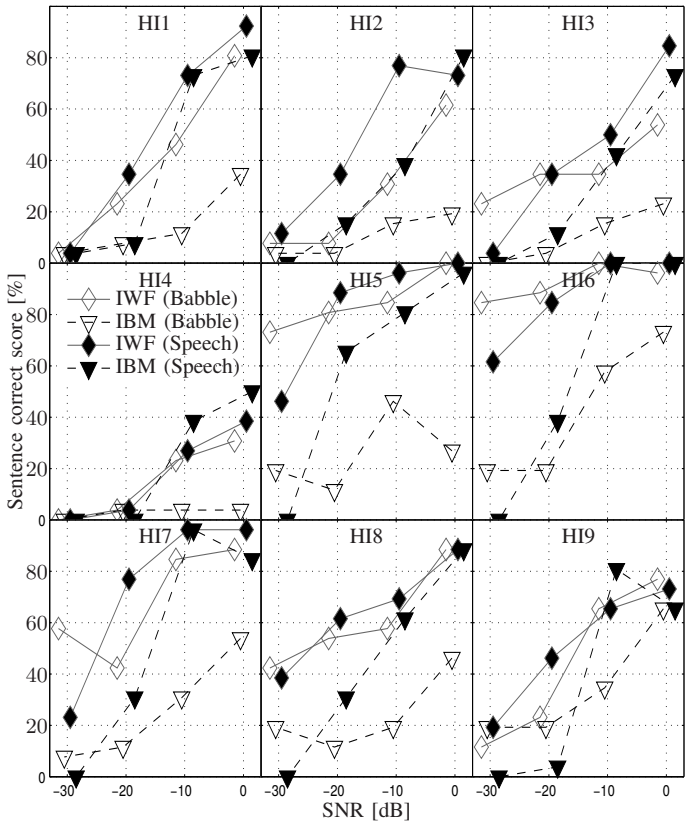

Fig. 2. Mean sentence correct scores averaged across sessions of the 9 HI listeners in the sentence recognition task with ideal parameter estimates with the VU sentences as the target speech material. The sentence correct scores for all subjects for the IWF and the IBM processed signals are marked by diamonds and triangles, respectively. The speech-in-babble and speech-in-speech conditions are denoted by empty and filled markers, respectively.

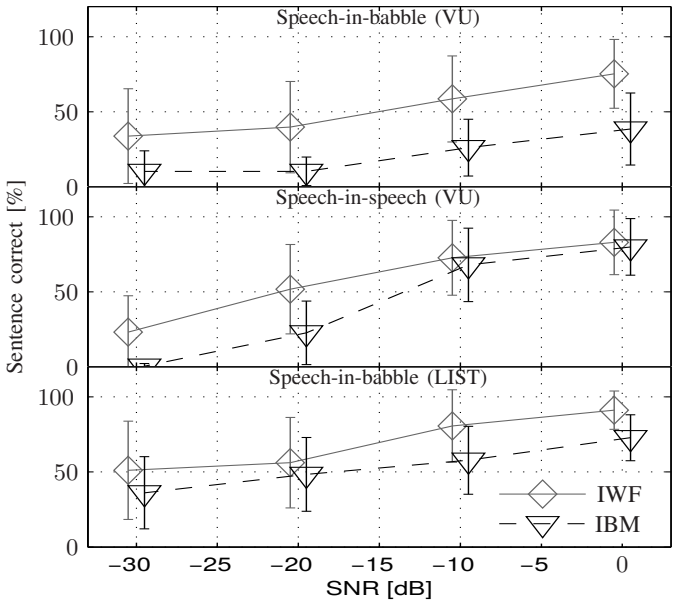

Fig. 3. Mean sentence recognition scores of all 9 HI listeners for the speech-in-babble and the speech-in-speech condition with the VU sentences are shown at the top and in the middle, respectively. The results for the speech-in-babble condition with the LIST sentences are shown at the bottom. The SI with the IWF and the IBM processed signals is shown by the grey solid line with diamonds and the black dashed line with triangles, respectively. The error bars depict one standard deviation of the scores.

RM-ANOVA with the factors session, noise scenario, NR algorithm and SNR. All factors passed Mauchly's test of sphericity. For all main factors, significant effects were obtained. The main factor of the factor session $[F(1,8) = 14.9;\ p < 0.05]$ corresponded to an overall increase of the scores in the retest session by 4.4%. The effect of the factor NR algorithm $[F(1,8) = 35.1;\ p < 0.05]$ corresponded to an overall increase of the SI with the IWF processed signals of $\Delta_{\mathrm{SI}} = 22.7\%$ compared to the IBM processed signals. Furthermore, the factors noise scenario $[F(1,8) = 41.2;\ p < 0.001]$ and SNR $[F(3,24) = 146.13;\ p < 0.001]$ were significant. A two-way interaction occurred between the factors noise scenario and SNR $[F(3,24) = 29.7;\ p < 0.001]$, and noise scenario and NR algorithm $[F(1,8) = 13.1;\ p < 0.05]$. A three-way interaction was also obtained between the factors noise scenario, SNR and NR algorithm $[F(3,24) = 7.4;\ p < 0.05]$. Furthermore, there was an interaction effect between all four factors obtained $[F(3,24) = 5.9;\ p < 0.05]$.

In the speech-in-babble scenario, *post-hoc* analysis of the scores at each SNR revealed that the SI was significantly different at SNRs of 0 dB ($p < 0.05$) ($\Delta_{\mathrm{SI}} = 36.8\%$), -10 dB ($p < 0.001$) ($\Delta_{\mathrm{SI}} = 32.5\%$), -20 dB ($p < 0.05$) ($\Delta_{\mathrm{SI}} = 29.5\%$), and -30 dB ($p < 0.05$) ($\Delta_{\mathrm{SI}} = 23.5\%$). At -30 dB, a significant effect of the interaction between the NR algorithm and the session was obtained. In the speech-in-speech condition, *post-hoc* analysis revealed that the difference in SI between the NR algorithm was significant at -20 dB ($p < 0.001$) and -30 dB ($p < 0.05$) SNR corresponding to a SI difference of $\Delta_{\mathrm{SI}} = 29.1\%$ and $\Delta_{\mathrm{SI}} = 22.6\%$, respectively.

For the sentence recognition task in the speech-in-babble scenario with the LIST sentences, the analysis of the scores was done with a three-way RM-ANOVA with the factors NR algorithm, SNR and session. Mauchley's test of sphericity was passed for all factors. The effect of the NR algorithm $[F(1,8) = 13.5;\ p < 0.05]$ corresponding to a SI difference of $\Delta_{\mathrm{SI}} = 16\%$ and of the SNR $[F(3,24) = 29.8;\ p < 0.001]$ was statistically significant. No significant effect of the factor session was obtained. The only significant interaction effect was between the factors session and mask $[F(1,8) = 5.9;\ p < 0.05]$. *Post-hoc* analysis revealed, that the interaction was based on a significant effect of the factor session on the SI with the IBM processed signals. The difference between the test and the retest session was 12.5%. Comparing the SI of both NR algorithms at fixed SNRs revealed that there were significant effects of the NR algorithm at 0 dB ($p < 0.05$) and -10 dB ($p < 0.05$) corresponding to differences of $\Delta_{\mathrm{SI}} = 18.3\%$ and $\Delta_{\mathrm{SI}} = 22.8\%$.

Furthermore, the correlation between the PTA of the subject (Listed in the final column of Table I) and the intelligibility scores (transformed to rationalized arcsine units) was investigated by means of a multiple regression analysis. The analysis was done with the factors NR algorithm, SNR, and session. For the VU sentences, the noise scenario was added as a factor. For the VU sentences, the factors NR algorithm, SNR, PTA and noise scenario contributed significantly and explained about 67% of the variance of the results. For the LIST sentences, the factors NR algorithm, SNR and PTA described the obtained results best and explained up to 57% of the variance.

For both speech corpora, the factor PTA was significant. Remarkably, the coefficient B which describes the contribution of a factor to the model was almost the same for both speech materials ($B_{\mathrm{VU}} = -1.63$ and $B_{\mathrm{LIST}} = -1.7$). Therefore, a higher degree of hearing loss represented by the PTA corresponds to lower SI scores with both NR algorithms. This indicates the the upper limit score that can be obtained with the studied NR algorithm depends on the hearing loss of the respective subject.

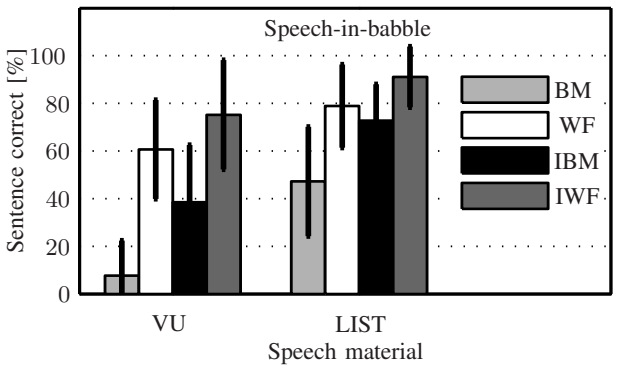

Fig. 4. Sentence recognition scores of the 9 HI listeners for the VU (left) and the LIST (right) sentences in the second sentence recognition task for the BM (light grey) and the WF (white) processed signals. For comparison, the scores with the IBM (dark grey) and the IWF (grey) processed signals with ideal parameter estimates are also shown. Error bars depict one standard deviation of the scores.

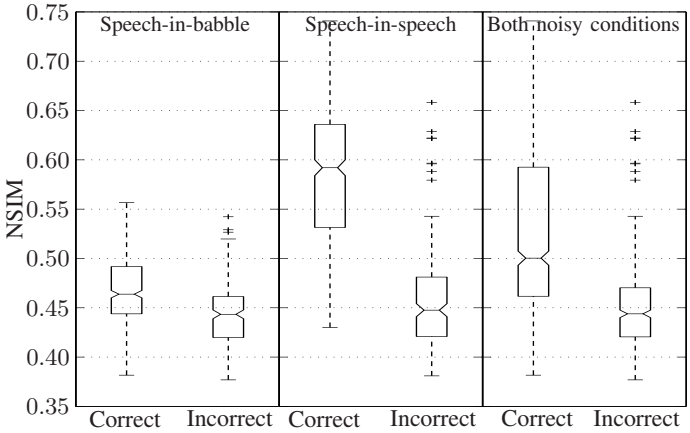

Fig. 6. NSIM value of the pooled data across noise reduction algorithm for the NH listeners for the speech-in-babble (left), speech-in-speech (middle) and the additionally pooled data across both interfering noises (right) for the intelligible (left boxplot) and the unintelligible sentences (right boxplot).

B. Robustness to estimation errors

The results of the speech recognition task with corrupted parameter estimates in the mask pattern derivation are shown in Fig. 4 for the group of HI listeners. The scores with the BM, WF, IBM and IWF processed signals are shown for the VU sentences (left) and the LIST sentences (right) as the light grey, white, black and dark grey bar, respectively. The error bars depict one standard deviation of the sentence correct scores.

The data were analysed with a four-way RM-ANOVA with the factors parameter estimate (ideal and corrupted estimates), speech material, mask pattern and session. Significant effects were obtained for the factors parameter estimate $[F\,(1,6) = 32.1;\ p < 0.05]$, speech material $[F\,(1,5) = 356.1;\ p < 0.001]$, and mask pattern $[F\,(1,6) = 94.1;\ p < 0.001]$. The effects corresponded to a SI difference of 21% between the ideal and corrupted parameter estimates, 27% between the factor speech material and $\Delta_{\mathrm{SI}} = 35\%$ between the factor mask pattern. Furthermore, significant interaction effects between the parameter estimate and the mask pattern $[F\,(1,6) = 13.2;\ p < 0.05]$ and the factor speech material and the mask pattern $[F\,(1,6) = 25.9;\ p < 0.05]$ were obtained. There was no other interaction effect obtained.

For the VU sentences, *post-hoc* analysis of the data revealed that, for both NR algorithms, the effect of the parameter estimate was significant. For the soft-decision and binary mask approach, the SI differences were 15% $(p < 0.05)$ and 31% $(p < 0.001)$, respectively. A comparison of the scores obtained with the corrupted parameter estimates showed a significant SI difference of $\Delta_{\mathrm{SI}} = 53\%$ $(p < 0.001)$ between the WF and the BM processed signals. For the LIST sentences *post-hoc* analysis showed that the sentence intelligibility of the WF and BM differed highly significantly $(p < 0.001)$from each other by $\Delta_{\mathrm{SI}} = 31.7\%$.

C. Preference rating

For each preference score, an absolute value ranging from 1 (imperceptible) to 5 (hugely better) was assigned with the sign according to the overall preference. Therefore, a positive value was assigned when condition A was preferred over B and a negative value if *vice versa*. The mean score of this values was determined for each subject for the statistical analysis. The test- and retest data were pooled for each subject. A nonparametric Wilcoxon signed rank test was conducted for each pairwise comparison and each interferer.

For the HI listeners, the results of the pairwise comparison preference rating task are shown in Fig. 5 for the speech-in-babble scenario for the VU (white bars) and the LIST (grey bars) sentences. The results for the VU sentences in the speech-in-speech scenario are shown with the black bars.

For the VU sentences, the statistical analysis revealed that the clean speech signal was preferred over the IBM $(p < 0.001)$ and the IWF $(p < 0.001)$ in the speech-in-babble and speech-in-speech scenario. The IWF and the IBM processed signals were not significant different in terms of speech quality preference with both interfering background sounds.

The clean speech signal was also significantly $(p < 0.001)$ preferred over the IBM and IWF processed signal when the LIST sentences served as the speech material. Again, there was no significant preference between the IWF and the IBM processed signals.

V. Results of the model-based evaluation

The speech intelligibility results of the HI listeners showed a huge variability and statistical analysis revealed a correlation between the intelligibility score and the PTA of the respective subject. Therefore, an attempt was made to evaluate the processing strategies with an auditory-periphery model to explain the speech intelligibility of a sentence based on its neural representation and how that representation is affected by background noise, processing by a NR algorithm, and hearing loss.

The NSIM values of all processed VU sentences in NH and HI listeners were calculated. For the NH listeners, the hearing loss was set to 0 dB HL for all frequencies. For the HI listeners, adjustments were made for each subject based on the amplification rule for the processed stimuli and the audiogram was used as input of the auditory nerve model. The NSIM values were then grouped to the intelligibility of each sentence into two groups (correctly and incorrectly repeated sentences). In Fig. 6, the NSIM values for the intelligible (left boxplot) and unintelligible (right boxplot) sentences are depicted with boxplots for the speech-in-babble (left), speech-in-speech (middle) and grouped for both scenarios (right). The data was pooled across the noise reduction algorithms.

In total, 2024 sentences were repeated correctly by the NH listeners while 472 were not repeated correctly. Statistical analysis revealed that for the pooled data across noise reduction algorithms and interfering sounds, there was a significant difference in mean NSIM value between the intelligible and unintelligible sentences $t\,(2494) = 21.5$, $(p < 0.001)$. The mean NSIM value for the correctly repeated

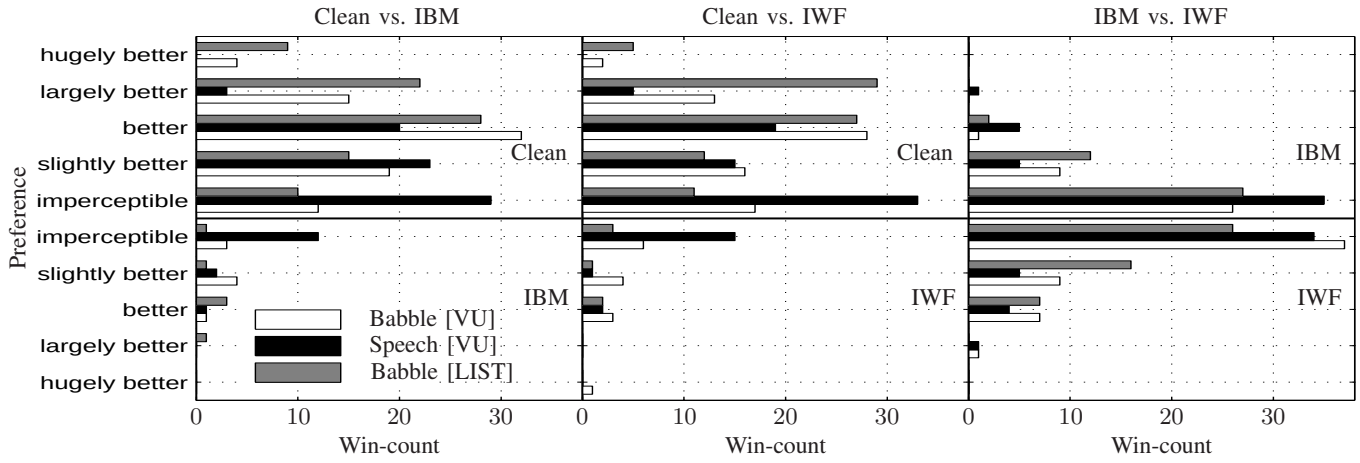

Fig. 5. Win-count histogram of the preference rating task and its amount of preference from the pairwise comparisons are shown for the HI listeners. The results for the VU sentences are shown with the white and the black bar for the speech-in-babble and the speech-in-speech scenario, respectively. The preference when the LIST sentences were processed in the speech-in-babble scenario are shown with the grey bars.

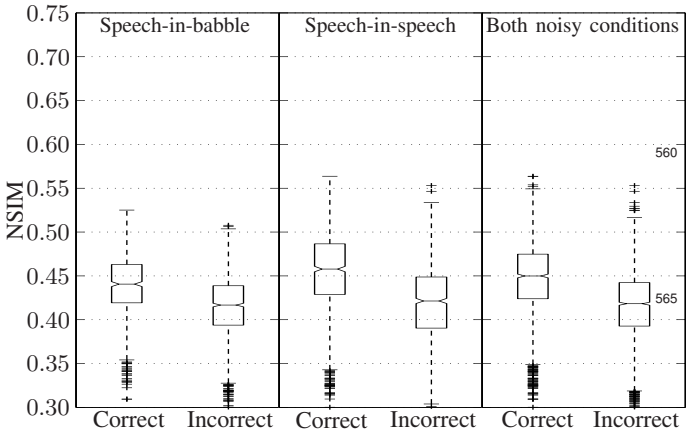

Fig. 7. NSIM value of the pooled data across noise reduction algorithm for the HI listeners for the speech-in-babble (left), speech-in-speech (middle) and the additionally pooled data across both interfering noises (right) for the intelligible (left boxplot) and the unintelligible sentences (right boxplot).

sentences was $\bar{x}_{\mathrm{cor}} = 0.495$ and for incorrectly repeated sentences was $\bar{x}_{\mathrm{incor}} = 0.447$. Analysis of the sentence intelligibility of the IWF processed signals revealed a highly significant difference ($t\,(1246) = 3.7,\ (p < 0.001)$) in the group means of the NSIM value of the intelligible and unintelligible sentence ($\bar{x}_{\mathrm{cor}} = 0.514$ and $\bar{x}_{\mathrm{incor}} = 0.490$). A significant difference in the NSIM value of the intelligible and unintelligible sentences for the IBM processed signals was obtained as well ($t\,(1246) = 14.4,\ (p < 0.001)$). The mean NSIM value of the correctly identified sentences $\bar{x}_{\mathrm{cor}} = 0.469$ differed significantly from the mean NSIM value of the non-identified sentences $\bar{x}_{\mathrm{incor}} = 0.442$.

As can be seen in Fig. 6, the NSIM values were much higher in the speech-in-speech scenario than in the speech-in-babble scenario. However, the mean NSIM value of the unidentified sentences was almost the same in both interfering sounds in NH listeners.

In Fig. 7, the NSIM values for the intelligible and unintelligible sentences are shown for the group of HI listeners for the speech-in-babble (left), speech-in-speech (middle) and pooled across both interfering speech-in-noise scenarios (right) as boxplots. Due to the simulation of the hearing impairment of the HI listeners, the NSIM values were generally lower for HI than for NH listeners. Furthermore, the difference between the mean NSIM score of the speech-in-babble and speech-in-speech scenario was greatly reduced.

Statistical analysis of the pooled data across the interfering background sounds and the processing condition of the HI listeners of a total sample size of 3474 sentences revealed that there was a significant difference ($t\,(3742) = 21.6,\ (p < 0.001)$) between the mean NSIM score of the intelligible ($\bar{x}_{\mathrm{cor}} = 0.446$) and the unintelligible sentences ($\bar{x}_{\mathrm{incor}} = 0.415$). When the data is pooled across interfering sounds, the mean NSIM scores of recognized and unintelligible sentences for the HI listeners were significantly different for the IWF ($t\,(1870) = 12.1,\ (p < 0.001)$) and the IBM ($t\,(1870) = 15.7,\ (p < 0.001)$) processed sentences corresponding to mean scores of $\bar{x}_{\mathrm{cor}} = 0.449$ and $\bar{x}_{\mathrm{incor}} = 0.415$, and $\bar{x}_{\mathrm{cor}} = 0.443$ and $\bar{x}_{\mathrm{incor}} = 0.408$ for the IWF and IBM processed sentences, respectively. Note that the mean values of the unintelligible sentences were very close to each other for the IWF and IBM as it was the case for the NH listeners.

VI. Discussion

The SI performance of the HI listeners for the IWF processed signal was significantly better than for the IBM processed signals in the speech-in-babble and the speech-in-speech scenario. In contrast to NH listeners reported in this study and in [31] when the VU sentences served as the target speech material, the scores for the IWF processed signals dropped with decreasing SNR and SI performance close to 100% was not obtained. In general, the scores of the HI listeners were lower for both types of processed signals than the scores for the IBM processed signals with NH listeners. But in both groups of listeners, the IWF significantly outperformed the IBM in terms of SI with ideal parameter estimates.

The results of the NH listeners presented in this study are in very good agreement with the results obtained in [31]. Due to the chosen test-retest design in this study, it could be shown that there was no learning effect that could suggest that the SI of IBM processed sentences would rise with enough training. The session effect was not significant for both speech materials. In HI listeners, there was a very small session effect corresponding to an increase of 4.4% in SI of the IBM processed signals in the retest session. However, this effect was very small and the fact that no session effect was obtained when the LIST sentences were used as the target speech material shows that it is unlikely that the SI performance for both NR algorithms would substantially increase with more training sessions.

Furthermore, there was a significant effect of the interfering sound on the SI of the HI listeners with the VU sentences. The SI is overall lower in the speech-in-babble scenario, especially for the IBM processed signals at SNRs of 0 dB and -10 dB. For the IBM processed signals, performance was very poor below -20 dB while the SI for the IWF processed signals was between 30% and 40%. Generally, the scores were much lower for HI than for NH listeners.

The SI performance of the HI listeners was very poor in the speech-in-babble scenario with the VU sentences. Therefore, the LIST sentences were also included. The LIST materials were developed for listeners with a profound sensorineural hearing loss and have a slower speaking rate. As expected in both groups of listeners, there was a significant SI improvement when the LIST sentences were used as the target speech material in comparison to the VU sentences. However, there was still a significant difference between the results for the NR algorithms suggesting that the IWF outperforms the IBM independent of the speaking rate. In comparison to the VU sentences, the overall difference in SI for the mask patterns became smaller with the slower speaking rate of the LIST sentence.

The results obtained with the corrupted mask patterns revealed the same ordering of NR algorithms as with ideal parameter estimates. The WF processed signals outperformed the BM processed signals in terms of SI and the performance for the WF processed signals was on average better than for the IBM. This result is also in agreement with [31] that performed the task with NH listeners. The effect is present for both speech materials with higher scores for the LIST sentences than for the VU sentences.

The decrease in SI performance was larger for the BM than for the WF processed signals. While the target information in a time-frequency point can be completely removed when an under-estimation of the short-term SNR occurs in the BM, it is only weighted in a disadvantageous way in the WF pattern and not fully removed. An over-estimation of the SNR could result in isolated time-frequency points that could be perceived as musical noise. However, it has been shown with NH listeners that musical noise introduced by isolated time-frequency points does not decrease SI of IBM processed signals [25, 27]. Therefore, the drop in SI of the BM processed signals can be most probably attributed to under-estimation errors in the mask pattern.

In [32], the WF processed signals outperformed the BM processed signals in terms of SI in NH subjects listening to noise vocoded signals but performance was lower than for the IBM and IWF processed signals. In CI users, there was no difference between the BM and IWF obtained. In comparison to CI users, the results suggest that HI listeners are sensitive to estimation errors in the mask pattern. The results suggest that the envelope processing and the electrical stimulation of a CI results in a reduced sensibility to speech distortions in terms of SI [49].

In [30], HI subjects listened to binary mask processed speech with mask patterns estimated from the noisy mixture. In our study, the binary mask pattern was calculated with perfect parameter estimates. The scores in [30] obtained at 0 dB SNR were close to 100 % for almost all HI listeners that had a PTA range from 33 dB to 54 dB HL. The scoring was done on component words. In this study, a sentence was scored as intelligible when all words were recognized correctly. The scoring level can explain why the results in this study were lower although the mask pattern was derived with ideal parameter estimates. Another reason is that the threshold for the IBM in this study was set to the input SNR. It has been shown that a threshold value lower than the input SNR is better for SI of BM processed speech than a threshold equal to the input SNR or a fixed threshold of 0)dB to maximize SNR gain [25, 30, 34]. Furthermore, the processing was done in [30] with a narrower frequency resolution on 64 frequency bands. As mentioned before, SI of IBM processed speech increases with increasing frequency resolution [27]. While in [30] the benefit was largest for HI subjects with high PTAs in comparison with unprocessed stimuli, we showed that the PTA of the subject could explain part of the variance of the obtained scores of both speech materials indicating that a higher PTA corresponds to lower scores with both NR algorithms.

For both speech corpora, the clean signals were preferred over the IBM and IWF processed signals in terms of speech quality. The HI listeners were able to differentiate the unprocessed signal from the processed signals and the influence of the interferer led to a decrease in perceived quality. Furthermore, the SI of a sentence could also influence the preference rating because not all HI subjects scored at 100 %, in particular with the IBM processing and the VU sentences as target speech material. However, the effect of SI was reduced because the subjects were allowed to listen as often as they wanted to the stimulus but an influence of SI on the preference can not be completely excluded.

Although a significant difference in SI was obtained between the NR algorithms at 0 dB SNR in the speech-in-babble scenario, there was no significant preference for one of the NR algorithms. Although NH listeners in [31] reported a clear preference for the IWF processed signals in both listening scenarios that was attributed to the higher amount of distortions introduced by the IBM processing, the HI listeners did not show this preference for either speech corpus or either noise scenario. Most probably, the spectro-temporal disjointness caused by the IBM processing is less prominent in HI listeners due to the reduced spectral and temporal resolution. Therefore, the better preserved envelope of the IWF processed signal was not perceived as distinct from the IBM processed signal in HI listeners as in NH subjects. The results of the preference rating are more in agreement with results obtained in CI users [32] that showed also no significant preference for one of the NR algorithms.

The results of the preference rating can be explained additionally by the reduced detection threshold of distortions introduced by the processing. [50] showed that the just noticable difference in the detection of speech and noise distortion is in HI listeners much higher than in NH listeners. Furthermore, [51] conducted a quality rating with NH and HI listeners of signals processed with different NR strategies and showed that the perceived speech naturalness differed between both groups.

In this study, a first attempt was made to evaluate time-frequency masking noise reduction algorithms with an instrumental measure that is based on the simulated auditory nerve response using a computational model. It was shown that the groups of intelligible and unintelligible sentences for both groups of listeners had distinct distributions of NSIM values. The mean NSIM score differed significantly for the pooled data across noise reduction algorithms for the intelligible and unintelligible sentences. While the mean scores were different for both speech-in-noise conditions for the intelligible sentences, the mean score of the unintelligible sentences was constant suggesting that it is possible to determine a threshold where sentences with lower NSIM are most probably not intelligible. If such a threshold could be derived for a larger set of noise reduction strategies and testing conditions for NH and HI listeners, the development of noise reduction strategies for the application in auditory prostheses (i.e. hearing aids) would hugely benefit from the use of the NSIM-measure. In a first development phase, different strategies could be evaluated with an auditory nerve model that could simulate a number of degrees of hearing losses of HI listeners.

While the NSIM scores were found to show significant differences, there was a fairly large spread of values around the means for correctly and incorrectly identified sentences (see Figs. 6 and 7).

This may be due in large part to the fact that the perceptual scores were based on complete identification of all words in a sentence, whereas the NSIM measure rewards partial matches between the neural representations of the test and reference sentences. A more sophisticated prediction scheme would be required to predict word confusions and thus generate estimates of complete sentence recognition. In addition, there are a number of different neural-based speech intelligibility predictors that could be explored as alternatives to the NSIM measure [52, 53, 54, 55]

The SI results obtained in HI listeners are in agreement with the results obtained with NH listeners in [31] in terms of the better SI provided by the soft mask with ideal parameter estimates and robustness of the SI performance with respect to estimation errors in the mask patterns. The soft-decision approach also outperformed the binary approach at very low SNRs. However, the SI was not close to 100 % with ideal parameter estimates suggesting that there is a limit in the benefit that HI listeners can obtain with NR strategies. The benefit appears to be dependent on the degree of hearing loss of the subject. The distortions introduced by both NR algorithms decreased SI in HI listeners. It is not surprising that perfect SI cannot be restored by the proposed NR algorithms because the pre-processing cannot bypass the problems that accompany a sensorineural hearing loss such as the previously mentioned reduced frequency resolution, reduced masking release and limited use of temporal fine structure.

The proposed mask patterns with the coarse spectral resolution of 22 Bark bands do not have the potential to restore perfect SI in HI listeners. Note that the coarse spectral resolution is representative or even higher than the spectral resolution for single channel NR pre-processing in state-of-the-art hearing aids. In HI listeners, a binary mask algorithm provided better scores than obtained in this study without the assumption of ideal parameter estimates [30]. However, the chosen SNR for HI listeners in [30] was very high and it is unclear what the limits of the processing are with ideal parameter estimates and at very low SNRs. For HI listeners, a higher spectral resolution of the time-frequency masking could lead to better SI scores because NH listeners benefited from an increased number of processing channels [27]. However, increasing the frequency resolution of the algorithm may not help if the HI listeners have impaired cochlear frequency resolution.

The result of the preference rating suggests that the NR applied in HAs can be more aggressively tuned for HI listeners without influencing the perceived quality of the processed signal. HI listeners seem to be less sensitive to distortions introduced by speech enhancement strategies. The variation from NH to HI listeners is in agreement with former results obtained in CI [32, 49].

VII. Conclusion

In conclusion, the IBM and IWF NR approaches were not able to restore perfect SI in HI listeners with the processing done on a spectral resolution of 22 critical bands according to the Bark scale. However, the soft-decision approach of the WF outperformed the binary approach of the BM in terms of SI with ideal and perturbed parameter estimates. In terms of speech quality, there was no preference for one of the NR algorithms. The further development of NR strategies in HAs should focus on higher spectral resolutions for time-frequency masking and can be tuned more aggressively due to the reduced sensitivity to distortions introduced by the processing.

References

[1] R. Plomp, "Noise, amplification, and compression: Considerations of three main issues in hearing aid design," *Ear Hear*, vol. 15, no. 1, pp. 2–12, 1994.

[2] L. S. Eisenberg, D. D. Dirks, and T. S. Bell, "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," *J Speech Hear Res*, vol. 38, no. 1, p. 222, 1995.

[3] G. A. Takahashi and S. P. Bacon, "Modulation detection, modulation masking, and speech understanding in noise in the elderly," *J Speech Lang Hear Res*, vol. 35, no. 6, pp. 1410–1421, 1992.

[4] R. Carhart and T. W. Tillman, "Interaction of competing speech signals with hearing losses," *Arch Otolaryngol Head Neck Surg*, vol. 91, no. 3, p. 273, 1970.

[5] T. Baer and B. C. Moore, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *J Acoust Soc Am*, vol. 95, no. 4, pp. 2277–2280, 1994.

[6] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J Acoust Soc Am*, vol. 88, no. 4, pp. 1725–1736, 1990.

[7] J. G. Bernstein and K. W. Grant, "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *J Acoust Soc Am*, vol. 125, no. 5, pp. 3358–3372, 2009.

[8] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B. C. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc Natl Acad Sci U S A*, vol. 103, no. 49, pp. 18866–18869, 2006.

[9] B. Moore, *An introduction to the psychology of hearing*. Academic press, San Diego, 5 ed., 2003.

[10] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans Audio Speech Lang Proc*, vol. 27, pp. 113–120, 1979.

[11] R. C. Hendriks and R. Martin, "Map estimators for speech enhancement under normal and rayleigh inverse gaussian distributions," *IEEE Trans Audio Speech Lang Proc*, vol. 15, pp. 918–927, 2007.

[12] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans Audio Speech Lang Proc*, vol. 14, pp. 1218–1234, 2006.

[13] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J Acoust Soc Am*, vol. 122, pp. 1777–1786, 2007.

[14] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun*, pp. 588–601, 2007.

[15] H. Levitt, "Noise reduction in hearing aids: a review.," *J Rehabil Res Dev*, vol. 38, no. 1, 2001.

[16] S. Gustafson, R. McCreery, B. Hoover, J. G. Kopun, and P. Stelmachowicz, "Listening effort and perceived clarity for normal-hearing children with the use of digital noise reduction," *Ear Hear*, vol. 35, no. 2, pp. 183–194, 2014.

[17] G. H. Mueller, J. Weber, and B. W. Y. Hornsby, "The effects of digital noise reduction on the acceptance of background noise," *J Am Acad Audiol*, vol. 24, pp. 649–659, 2013.

[18] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *J Acoust Soc Am*, vol. 127, pp. 2054–2063, 3 2010.

[19] M. Boymans and W. A. Dreschler, "Field trials using a digital hearing aid with active noise reduction and dual-microphone directionality: Estudios de campo utilizando un audifono digital con reduccion activa del ruido y micrófono de direccionalidad

dual," *Int J Audiol*, vol. 39, no. 5, pp. 260–268, 2000.

[20] J. Alcántara, B. Moore, V. Kühnel, and S. Launer, "Evaluation of the noise reduction system in a commercial digital hearing aid.," *Int J Audiol*, vol. 42, no. 1, p. 34, 2003.

[21] P. Vary and R. Martin, *Digital Speech Transmission - Enhancement, Coding and Error Concealment*, ch. Single and Dual Channel Noise Reduction, pp. 389–466. John Wiley & Sons, Ltd., Chichester, 2006.

[22] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer, Berlin, 1 ed., 2005.

[23] D. Wang, *Speech Seperation by Humans and Machines*, ch. On ideal binary mask as the computational goal of auditory scene analysis, pp. 181–197. Norwell, MA: Kluwer, 2005.

[24] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Commun*, vol. 51, no. 3, pp. 230–239, 2009.

[25] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency," *J Acoust Soc Am*, vol. 120, pp. 4007–4018, 2006.

[26] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear Hear*, vol. 27, no. 5, p. 480, 2006.

[27] N. L. Li and P. C. Loizou, "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *J Acoust Soc Am*, vol. 123, pp. EL59–EL64, 3 2008.

[28] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J Acoust Soc Am*, vol. 125, no. 4, pp. 2236–2347, 2009.

[29] G. Kim, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J Acoust Soc Am*, vol. 126, no. 3, pp. 1486–1494, 2009.

[30] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J Acoust Soc Am*, vol. 134, no. 4, pp. 3029–3038, 2013.

[31] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Trans Audio Speech Lang Proc*, vol. 21, pp. 63–72, 1 2013.

[32] R. Koning, N. Madhu, and J. Wouters, "Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," *IEEE Trans Biomed Eng*, vol. 62, no. 1, pp. 331–341, 2015.

[33] M. S. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics," *J Acoust Soc Am*, vol. 126, pp. 2390–2412, 2009.

[34] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J Acoust Soc Am*, vol. 126, pp. 1415–1426, 2009.

[35] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans Audio Speech Lang Proc*, vol. 19, pp. 47–56, 1 2011.

[36] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J Acoust Soc Am*, vol. 33, no. 2, p. 248, 1961.

[37] H. Dillon, *Hearing aids*. Thieme Publishers New York, 2012.

[38] N. J. Versfeld, L. Daalder, J. M. Festen, and T. Houtgast, "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," *J Acoust Soc Am*, vol. 107, pp. 1671–1684, 2000.

[39] A. van Wieringen and J. Wouters, "LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for flanders and the netherlands," *Int J Audiol*, vol. 47, pp. 348–355, 2008.

[40] M. S. Zilany and I. C. Bruce, "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *J Acoust Soc Am*, vol. 120, no. 3, pp. 1446–1466, 2006.

[41] M. S. Zilany and L. H. Carney, "Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics," *J Neurosci*, vol. 30, no. 31, pp. 10380–10390, 2010.

[42] M. C. Liberman, "Auditory-nerve response from cats raised in a low-noise chamber," *J Acoust Soc Am*, vol. 63, no. 2, pp. 442–455, 1978.

[43] C. J. Plack, V. Drga, and E. A. Lopez-Poveda, "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss," *J Acoust Soc Am*, vol. 115, no. 4, pp. 1684–1695, 2004.

[44] B. C. Moore, "Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants," *Otol Neurotol*, vol. 24, no. 2, pp. 243–254, 2003.

[45] E. A. Lopez-Poveda and P. T. Johannesen, "Behavioral estimates of the contribution of inner and outer hair cell dysfunction to individualized audiometric loss," *J Assoc Res Otolaryngol*, vol. 13, no. 4, pp. 485–504, 2012.

[46] A. Hines and N. Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Commun*, vol. 54, no. 2, pp. 306–320, 2012.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Image Proc*, vol. 13, no. 4, pp. 600–612, 2004.

[48] G. A. Studebaker, "A "rationalized" arcsine transform," *J Speech Hear Res*, vol. 28, pp. 455–462, 1985.

[49] O. Qazi, B. van Dijk, M. Moonen, and J. Wouters, "Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility," *Hear Res*, vol. 299, pp. 78–87, 2013.

[50] I. Brons, W. A. Dreschler, and R. Houben, "Detection threshold for sound distortion resulting from noise reduction in normal-hearing and hearing-impaired listeners," *J Acoust Soc Am*, vol. 136, pp. 1375–1384, 9 2014.

[51] M. Marzinzik, *Noise reduction schemes for digital hearing aids and their use for the hearing impaired*. PhD thesis, University Oldenburg, Germany, 2001.

[52] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech Commun*, vol. 41, no. 2, pp. 331–348, 2003.

[53] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Commun*, vol. 65, pp. 75–93, 2014.

[54] J. Swaminathan and M. G. Heinz, "Psychophysiological analyses demonstrate the importance of neural envelope coding for speech perception in noise," *J Neurosci*, vol. 32, no. 5, pp. 1747–1756, 2012.

[55] M. E. Hossain, W. A. Jassim, and M. S. Zilany, "Reference/free assessment of speech intelligibility using bispectrum of an auditory neurogram," *PLoS ONE*, vol. 11, no. 3, p. e0150415, 2016.